DEFENCE
SCIENCE & TECHNOLOGY

DSTO

The Effects of Background Noise on the Performance of an Automatic Speech Recogniser

Jason Littlefield and
Ahmad Hashemi-Sakhtsari

DSTO-RR-0248

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

20030610 068

DEFENCE
SCIENCE & TECHNOLOGY

The Effects of Background Noise on the Performance of an Automatic Speech Recogniser

Jason Littlefield and Ahmad Hashemi-Sakhtsari

Command and Control Division
Information Sciences Laboratory

DSTO-RR-0248

ABSTRACT

Ambient or environmental noise is a major factor that affects the performance of an automatic speech recogniser. Large vocabulary, speaker-dependent, continuous speech recognisers are commercially available. Speech recognisers perform well in a quiet environment, but poorly in a noisy environment. Speaker-dependent speech recognisers require training prior to them being tested, where the level of background noise in both phases affects the performance of the recogniser. This study aims to determine whether the best performance of a speech recogniser occurs when the levels of background noise during the training and test phases are the same, and how the performance is affected when the levels of background noise during the training and test phases are different. The relationship between the performance of the speech recogniser and upgrading the computer speed and amount of memory as well as software version was also investigated.

RELEASE LIMITATION

Approved for public release

AQ F03-07-1433

Published by

DSTO Information Sciences Laboratory
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

© Commonwealth of Australia 2002
AR-012-500
November 2002

APPROVED FOR PUBLIC RELEASE

The Effects of Background Noise on the Performance of an Automatic Speech Recogniser

Executive Summary

Large vocabulary, speaker-dependent, continuous speech recognisers have become commercially available in recent years and perform well in quiet environments. However, ambient or environmental noise is a major factor that affects the performance of speech recognisers. Potential areas of applications of these systems in the military include data entry in Command, Control, Communication and Intelligence (C3I) environments, technology-based training, and collaborative meetings or interviews. A prominent type of background noise in C3I environments is human conversation. It is important to identify the limitations of this emerging technology before developing applications in C3I environments.

Speaker-dependent speech recognisers require a training phase prior to them being used for testing. The training phase builds a user profile by combining a vocabulary and a regional language model with the phonetic analysis of the speaker's voice. The test phase transcribes the speaker's voice into text with the aid of the user profile produced during the training phase.

This research had two main objectives. Firstly, to determine whether the best performance of an automatic speech recogniser occurs when the levels of background noise during the training and testing phases are the same. Secondly, to ascertain how the performance is affected when the levels of background noise during training and testing are different.

Signal-to-noise ratio (SNR) was used to measure the difference between the speech and the background noise levels in decibels (dB). The automatic speech recogniser used in this experiment was Dragon Systems' Dragon NaturallySpeaking (NS) Professional version 4.0 and the scoring program used to measure the performance of NS in terms of percentage word recognition accuracy was Sclite from the US National Institute of Standards and Technology.

The experiment consisted of six major steps. Initially speech from 10 different speakers and one background conversation were recorded, digitised and saved as files. Each speaker audio file was then mixed with the background conversation at five different SNRs of 15dB, 20dB, 30dB, 40dB and 50dB. Using a computer equipped with a 400MHz processor and 128MB of memory, NS was trained with each of these mixed audio files to produce five user profiles for each speaker. NS was subsequently tested

using each of the combinations of the five mixed audio files and five user profiles to produce 25 hypothesis text files for each speaker. Transcribing the original speaker audio files by hand produced reference text files for all speakers. The reference and hypothesis texts were then compared using Sclite to produce 25 reports for each speaker, which includes the percentage word recognition accuracy.

This study revealed that there were two independent factors that affect the performance of NS in noisy environments: the difference between the level of background noise in the training and test environments, and the level of background noise in the test environment alone. The significance of these factors leads to a conclusion that regardless of the training environment, testing with the least amount of background noise achieves the best performance in terms of word recognition accuracy. However, for a particular test environment, the best performance was achieved when the levels of background noise during training and testing were the same. Additionally, the effect of the level of background noise in the test environment (with SNR in the range 15-50dB) accounts for variation in performance of up to 25%, and the difference between the level of background noise in the test and training environment accounts for variation of up to 4%. If the SNR for the test environment is greater than 30 decibels there should be little degradation in performance of NS due to noise, and NS could not be trained when the SNR was less than 15 dB.

The SNR in the test environment was found to affect the computer processor utilisation while transcribing, where decreasing the Test SNR increased the average computer processor utilisation. Another factor that affects the average computer processor utilisation is the computer processor speed. Dragon Systems claim each new version of NS provides improved performance. This improvement is achieved through better acoustic models, language models, and search algorithms. Consequently an experiment was conducted to quantitatively assess the effects of upgrading the computer processor speed and NS software on the word recognition accuracy. Increasing the computer processor speed from 400MHz to 1000MHz and the amount of physical memory (RAM) from 128MB to 512MB produced an increase in performance of about 10%. Additionally, upgrading NS from version 4.0 to 5.0 increased the performance by about 4%.

Many other factors were observed to affect the performance of NS, such as inherent differences between speakers, speaking rate, the degree of speaker enunciation and the stress level the speaker experienced. The combination of these factors resulted in variations in performance between speakers of up to 25%.

Authors

Jason Littlefield
Command and Control Division

Jason Littlefield graduated with a B.Sc. in the School of Mathematics and Computer Science from the University of Adelaide in 2000. He recently joined the Human Systems Integration Group as a professional officer in 2002. His areas of interest include Automatic speech recognition, Speech user interfaces and Speaker separation.

Ahmad Hashemi-Sakhtsari
Command and Control Division

Ahmad Hashemi-Sakhtsari is a research scientist in Human Systems Integration Group. His current research work is focused on application of commercial language technology to military systems and on studying human computer interaction through speech as well as manual modalities. He manages a small speech and language technology research and development task in the Human Systems Integration Group.

Contents

1. Introduction

Recent improvements in acoustic and language model accuracy, search algorithms and computing power have seen commercial-off-the-shelf (COTS) large vocabulary automatic speech recognisers (ASRs) achieve a high level of accuracy with continuous speech [Makhoul and Schwartz, 1994]. Due to the advances in technology that support efficient and natural means of user interaction with computers, ASR software has seen a large increase in popularity over the past few years. Verbal interaction with computers is useful when users' hands and eyes are in use or natural language interaction is preferred, especially when keyboard use is limited [Cohen and Oviatt, 1994].

ASRs can be categorised along different dimensions, these include speaker dependence, vocabulary size and speech continuity. Speaker dependence refers to the ASR being described as speaker-dependent, speaker-adaptive or speaker-independent. The vocabulary size refers to the number of words a user can speak and be recognised. Speaker continuity refers to whether words can be spoken in isolation, as connected speech or as continuous speech [Cohen and Oviatt, 1994].

Speaker-dependent ASRs require a training phase prior to them being tested. The training phase builds a user profile by combining a vocabulary and a regional language model with the phonetic analysis of the speaker's voice. The test phase transcribes the speaker's voice into text with the aid of the speaker profile produced in the training phase. The potential use of ASRs in a wide range of situations requires the systems to perform well with varying levels of background noise. Ambient and environmental noise, such as background conversation, causes a reduction in the performance of the ASR [Juang, 1991; Atal, 1994]. However, the reduction in the performance of the ASR may not be solely attributed to its use in a noisy test environment, but also a mismatch in the level of background noise between the training and test environments [Gong, 1991; Mammone and Zhang, 1998].

The primary goal of this research was to determine whether the optimum performance of an ASR occurs when the training and test environments have the same level of background noise. Additionally, the secondary goal was to investigate how training and testing with different levels of background noise affects the performance of an ASR. The motivation for this research comes from the demand for robust ASR applications in adverse military environments [Oberteuffer, 1994]. Identifying the limitations of current commercial ASRs is an important step toward providing suitable systems for particular military applications. One factor in the military's favour is that the size of the vocabulary in military environments is often limited [Weinstein, 1994]. Lippmann (1997) identified that the performance of ASRs improves as the size of the vocabulary decreases.

One potentially pervasive area of application of ASR is data entry in command and control centres [Weinstein, 1994]. In this situation the principal type of background noise is human conversation, and for this reason the type of background noise chosen for this experiment was also human conversation. The ASR software chosen for use throughout this project was Dragon Systems' Dragon NaturallySpeaking (NS) because it is a large vocabulary, speaker-dependent, COTS continuous speech recogniser that is widely reported to be accurate [Alwang, 1999; Plutchik, 2000]. To measure the performance of NS in terms of percentage word recognition accuracy, a software tool was selected from the US National Institute of Standards and Technology called Sclite (Score-lite), which is part of their Speech Recognition Scoring Toolkit.

In order to account for relatively low performance during the initial experiment, a complementary experiment was conducted to provide a broader depiction of factors that affect the performance of ASRs. The study analysed the effects of improvements in computing power and upgrading software version on the performance of ASRs.

The method for the experiment is described in section 2 of this report, while the results are discussed in section 3. Section 4 of this report provides the details and findings of this additional study. Section 5 provides further discussion of the findings and section 6 gives concluding remarks. The acknowledgements are mentioned in section 7, while the references are presented in sections 8. Appendix A and B outline the statistical analysis for the background noise, and computing power and software upgrade experiments respectively. Appendix C discusses the issue of making inferences from curves based on group data.

2. Background Noise Experimental Method

2.1 Overview

The audio signal from ten speakers and a background conversation (noise) were recorded, transferred to computer, digitised and saved as files. Each of the *speaker audio files* were mixed with the *noise audio file* at five different levels to produce five *mixed audio files*. NS was trained with each of the five *mixed audio files* from each speaker to produce five distinct *user profiles* for each Speaker. Following this, each of the five *user profiles* was used, one at a time, while testing NS with each of the five *mixed audio files* to produce 25 text files for each speaker. These text files are referred to as *hypothesis text files.* The speaker's voice from the original *speaker audio file* was transcribed by hand to produce a *reference text file.* The reference and hypothesis text files were compared using the Sclite program to produce *reports,* which include the percentage of words transcribed correctly.

2.2 Speaker Audio Files

Ten *speaker audio files* were produced using ten distinct speakers, with each *speaker audio file* consisting, on average, of about 4600 words spoken in English. Each *speaker audio file* was recorded using a microphone and a Sony DAT recorder while the speaker read aloud several chapters from the book 3001: The Final Odyssey [Clarke, 1997].

After the audio was recorded on the Sony DAT, they were transferred to a computer digitised and saved as files with a sampling rate of 22050 Hz, as single channel (mono) and resolution of 16 bits. The quality of audio recordings saved with this format is high enough for ASRs to perform well. The length of the recordings varied amongst the speakers from between 30 and 40 minutes. This difference was due to variation in the speech rate and number of times phrases were repeated.

A speaker's speech level and the background noise level in an operating environment can be measured using a Sound Pressure Level (SPL) meter (such as the Castle GA208 Sound Level Meter). The SPL meter can measure sound levels using different contour filters, the most commonly used being dBA and dBC using A and C contour filters respectively. The A-contour filter eliminates inaudible low frequencies and is designed to approximate what the human ear is capable of hearing [Nave, 2000]. The sound pressure level (SPL) of the recording room was determined to be 42dBA using an SPL meter.

2.3 Noise Audio File

One *noise audio file* was produced after mixing three individual recordings of human speech from two television programs and an audio compact disk. The two television programs recorded were an Australian current affairs program and a public speech by an Australian politician. The individual recordings were digitised and saved as audio files with a sampling rate of 22050 Hz, as single channel (mono) and with resolution of 16 bits. To mix the recordings together so that the audio level for each was the same, a 'C' program called Addwav was developed in-house. The Addwav program required a scaling factor to determine the level at which to mix the audio files. The audio level of audio recordings can be determined by calculating the Root Mean Sqaure (RMS) power. The RMS power of each of the three recordings was determined using another 'C' program called Wavpower also developed in-house. The Wavpower program uses equation 2.1 shown below to calculate the RMS power P of an audio file, where v_i represents the voltage on a sample period and N represents the number of samples periods [Kosbar, 1998].

$$P = \frac{1}{N}\sum_{i=1}^{N} v_i^2 \tag{2.1}$$

Two of the recordings had a smaller RMS power than the third recording. All three recordings were mixed together using the Addwav program with appropriate scaling factors so that the audio level for each was the same to produce a single *noise audio file.*

2.4 Mixed Audio Files

Fifty audio files were produced after mixing the speaker and noise audio files at five different levels for each of the ten speakers. The SNR was used to determine the level of noise to mix. The five different SNRs chosen were 10dB, 20dB, 30dB, 40dB and 50dB, where a SNR of 10dB corresponds to the speech level being 10dB higher than that of the noise. Hence, an audio file with a SNR of 10dB has the highest level of noise, i.e. a noisy environment, and a sound file with a SNR of 50dB has the lowest level of noise, i.e. a quiet environment. However, NS failed to train using an audio file with a SNR lower than 15dB. Consequently, 15dB SNR was chosen for the lower bound of the SNRs instead of 10dB. For each of the speaker audio files and the noise sound file, the RMS power was determined using the software program described in section 2.3. Each speaker audio file was mixed with the noise audio file five times, once for each of the SNR of 15dB, 20dB, 30dB, 40dB and 50dB.

The speaker and noise audio files were mixed together using the Addwav program. As before, the Addwav program required a scaling factor to determine the level at which to mix the speaker audio file and noise audio file. The Wavpower program was used again to determine the RMS power of the audio files. This time instead of mixing the audio files with the same audio level, they were mixed with SNRs of 15dB, 20dB, 30dB, 40dB and 50dB. Equation 2.2 describes the SNR in terms of the RMS power of the speaker (signal) audio file P_s and the RMS power of the noise audio file P_N [Lathi, 1998].

$$SNR = 10\log_{10}\left(\frac{P_S}{P_N}\right) \tag{2.2}$$

In all cases, the RMS power for the noise audio file was higher than the RMS power of the speaker audio file. For each speaker five scaling factors were calculated, one for each SNR of 15dB, 20dB, 30dB, 40dB and 50dB. Using the five scaling factors for each speaker, the audio files were mixed to produce five *mixed audio files* for each speaker, with SNRs of 15dB, 20dB, 30dB, 40dB and 50dB.

2.5 User Profiles for Dragon NaturallySpeaking

Fifty NS user profiles were produced, one for each of the fifty mixed audio files. Two personal computers were used in the training process to generate the user profiles. One computer was used to play the mixed audio files while another was used to generate a user profile using NS Professional version 4.0. The computer with NS was equipped with an Intel Pentium II® 400MHz processor, 128MB of memory and a SoundBlaster 64 sound card. The audio signal from the audio file was transferred via a stereo mini-jack lead from the sound card line output of the computer playing the audio file to the sound card line input of the computer with NS. About 600 words from each of the mixed audio files were used to generate the *User Profiles* using NS Professional version 4.0. The options selected during training NS version 4.0 were the

"UK BestMatch" speech model and the "UK General English–BestMatch Plus" vocabulary. Out of the 4600 words in the mixed audio files, 600 were used for training leaving the remainder (4000 words) for testing NS.

2.6 Reference Text Files

As described earlier, as part of the training process for NS, paragraphs from the text selection were displayed to the user, one at a time. NS does not display further paragraphs from the text selection until it has registered utterances from the speaker reading aloud the current paragraph. This may require several attempts at some words or utterances and only one for others. Hence, the length of each of the speaker audio files that were recorded while training NS from the same text selection differed. Thus, it was necessary to make a record of what was actually read aloud for each speaker. Each of the 10 Reference text files were produced by hand, with an experimenter transcribing each of the speaker audio files as they were played.

2.7 Hypothesis Text Files

NS was used to produce 250 hypothesis text files in total. As for the training process, two computers were used during the test process. On the computer with NS, a user profile for one of the speakers was loaded. NS was then used to transcribe each of the five mixed audio files for that speaker. This was repeated for each of the user profiles for that speaker, and then repeated for each of the speakers. This produced 25 hypothesis text files for each of the ten speakers. Not all of the mixed sound files were used in creating the hypothesis text files (4000 from the 4600 words); the sections that were used in the training process (600 words) were skipped.

2.8 Comparing Reference and Hypothesis Text Files

A scoring program called *Sclite* was used to compare the reference and hypothesis text files. Sclite is part of the Speech Recognition Scoring Toolkit (SCTK) version 1.2 from the US National Institute of Standards and Technology (NIST). Sclite was designed to compare text output from a speech recogniser such as NS (hypothesis text) to the original text (reference text) and generate a report summarising the performance. The comparing of the reference to the hypothesis text is called the alignment process. The gathering of statistics for the report is called the scoring process [Fiscus, 1998].

The alignment process consists of two steps: selection matching the reference and hypothesis text files, and performing a text alignment. From four algorithms available for selection matching, the Utterance Id matching method was used, which requires the reference and hypothesis files to be in transcript format (trn). A file in transcript format has word sequence records separated by new-lines, where each word sequence record consists of a series of words and blank spaces followed by an Utterance Id in parentheses. Using this format there were several Utterance Id types available and the

spu_id type was selected, where the spu_id Utterance Id consists of a speaker name followed by a hyphen and a number [Fiscus, 1998].

Conversion of the text files to transcript format was done by hand, with eight word sequence records in each file, for all of the 260 text files. The location of the Utterance Ids in each of the hypothesis text files was made to coincide with the same location in the reference text file (see example below). The selection matching locates corresponding reference and hypothesis texts with the same Utterance Ids.

Example:

Reference file
Professor Anderson was a small dapper man whose features seemed to have combined key aspects of several races Chinese Polynesian Nordic in a thoroughly confusing fashion (speaker1-1) [Clarke, 1997].

Hypothesis file
Professor and so was a small that the man whose features seem to have combined key aspects of several rises shining Polynesia Nordic in a thoroughly confusing fashion (speaker1-1).

Within *Sclite*, scores are compiled after the alignment process for each reference-hypothesis record pair, and a report is then generated. There are various types of reports available. The type of report can be nominated in the options *(-o)* part of the Sclite program. The report selected for this project was the *dtl* report, which provides the WRA as a percentage. Various error categories are also provided.

The size of the word sequence records in each file affects the synchronisation of the word strings. The smaller the word sequence the more accurate the scoring [Kemp et al., 2000]. However, an investigation into the effect of the length of the word sequence records on WRA using *Sclite*, showed that the WRA improved by only 1% when 128 word sequence records were used instead of eight (80.6% for 8 versus 81.6% for 128). In the experiment, each file of about 4600 words was broken up into eight records of about 575 words each. Since the task of converting the text files to transcript format was very time consuming and there were a large number of text files to convert, the number of word sequence records for each file was restricted to eight.

3. Background Noise Experimental Results

Percentage word recognition accuracy (WRA) was used as the performance measure of the ASR and was defined as the percentage of words recognised correctly for each transcription. Figure 1 shows a graph of the arithmetic means for the WRA for different Train and Test SNRs. There are two important factors that can be observed from the graph. Firstly, the effect of the Test SNR, within each Train SNR category, that appears to be logarithmic in nature, where the WRA degrades moderately from 50dB to 30dB, and dramatically from 30dB to 15dB. Secondly, the highest WRA for

each Test SNR category (i.e. vertically on the graph) occurred when the Test SNR matched the Train SNR.

A three-dimensional graph shown in figure 2 emphasizes how WRA is affected when the Train SNR and Test SNR match and when they mismatch.

The influence of a factor, such as the Test SNR, on a continuous dependent variable, such as the WRA, can be studied using Analysis of Variance (ANOVA) [Keppel, 1982]. ANOVA was conducted examining the influence of the Test SNR, the Train SNR and their interaction on the WRA. The Test SNR was determined to be a significant factor. Furthermore, when the Train SNR was treated as the absolute difference from the Test SNR, such as in equation 3.1, the SNR Difference was also found to be a significant factor.

$$SNR\ Difference = |Train\ SNR - Test\ SNR| \tag{3.1}$$

Multiple regression studies the relationship between a dependent variable and significant factors. This relationship can be described using a population regression equation. Equation 3.2, the population regression equation for the relationship between the WRA and the Test SNR and SNR Difference, accounts for 90.8% of the sampled data. Values close to 100% indicate that the model fits the data well [Moore and McCabe, 1993].

$$WRA = 5.65 + 31.22\ \log_{10}(Test\ SNR) - 0.13\ SNR\ Difference \tag{3.2}$$

Figure 3 shows variation in the WRA across different speakers. When trained and tested in a quiet environment, i.e. 50dB SNR, the best result for WRA was 71.6% and the lowest was 47.1%, with an arithmetic mean of 56.0%. Likewise, when trained and tested in a noisy environment, i.e. 15dB SNR, the best result for Words accuracy was 48.1% and the lowest was 29.2%, with an arithmetic mean of 40.3%. Hence the variation in the WRA between speakers was as much as nearly 25%, which is significant. However, the variation was, on average, about 15%.

The report produced by the scoring program Sclite gives the types of errors made by an ASR as well as the percentage of words transcribed correctly. The errors are divided into three types: substitution, deletion and insertion errors. A substitution error occurs when an ASR incorrectly transcribes a word, whereas a deletion error corresponds to a word being omitted. Deletion errors can be due to an input utterance having too low an amplitude or too short a duration, or because ASR has failed to match it to any of the reference patterns in its vocabulary. Insertion errors occur when extraneous words are inserted in the transcribed hypothesis texts that do not correspond to those in the original reference text. Out of the three types of errors, substitution errors were the most significant, accounting for about 60% of all errors, while the deletion and insertion accounted for approximately 38% and 2% respectively.

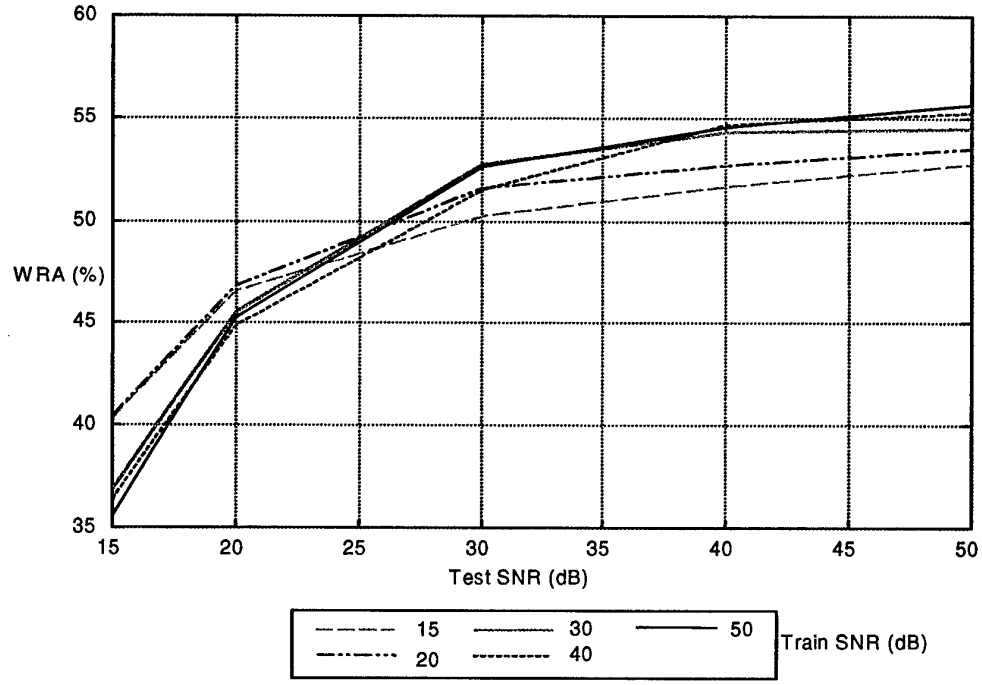

Figure 1 *Graph of WRA for different combinations of Train and Test SNRs. For each Train SNR category the highest WRA was achieved when the Test SNR was highest (i.e. the least noisy). At each Test SNR category the highest WRA occurred when the Train SNR matched the Test SNR.*

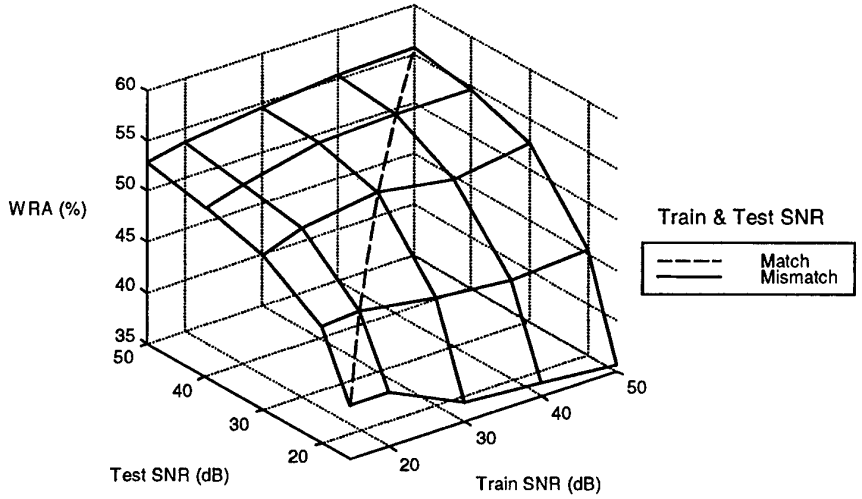

Figure 2 *Graph of WRA for different combinations of Train and Test SNRs, highlighting the performance when the Train and Test SNRs match and when they mismatch.*

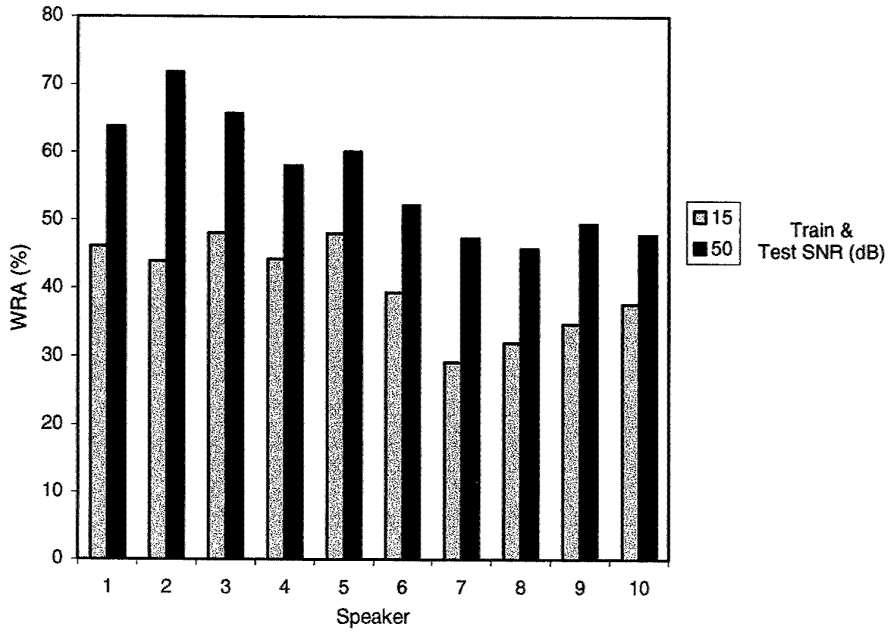

Figure 3 Graph of the WRA (%) for different Speakers when trained and tested in quiet (50dB SNR) versus noisy (15dB SNR) environments. Note the large variation between speakers.

ASRs make a trade off between speed and accuracy. The speed and accuracy is dependent on the resources available, i.e. the processor speed and amount of memory available. NS allows the user to control the speed versus accuracy trade off for speech recognition as one of its options with the default set to faster and less accurate transcribing.

Recall that the experiment was conducted using speaker audio files with duration of about 30 minutes and the computer used was equipped with a 400MHz processor and 128MB of memory. During testing, some of the audio files took longer than 30 minutes to transcribe, particularly the audio files with low SNR. The processor utilisation and memory usage during testing were measured using the Microsoft® WindowsNT® Workstation 4.0 (service pack 6a) performance monitor. The results are displayed in figure 4 where Speaker-3's 50dB SNR user profile was used for transcribing Speaker-3's 15dB, 20dB, 30dB and 50dB SNR mixed audio files. Analysis of the processor utilisation and memory usage over time during testing reveals the cause of the time lag for transcription.

The average processor utilisation for transcribing the 15dB, 20dB, 30dB and 50dB mixed audio files were 96.6%, 71.1%, 48.9% and 44.0% respectively. Note that, from observation, the variability in the processor utilisation over time was related to the rate of speech, where a sharp increase in processor utilisation corresponded to a long quickly spoken utterance, and a sharp decrease in processor utilisation corresponded

to a pause between spoken utterances. The time lag between the duration of the 15dB, 20dB, 30dB and 50dB mixed audio files and the time taken to transcribe were 7min, 1min, 0min and 0min respectively. Also note that the WRAs for the same 15dB, 20dB, 30dB and 50dB mixed audio files were 38.9%, 52.7%, 61.0% and 63.5% respectively.

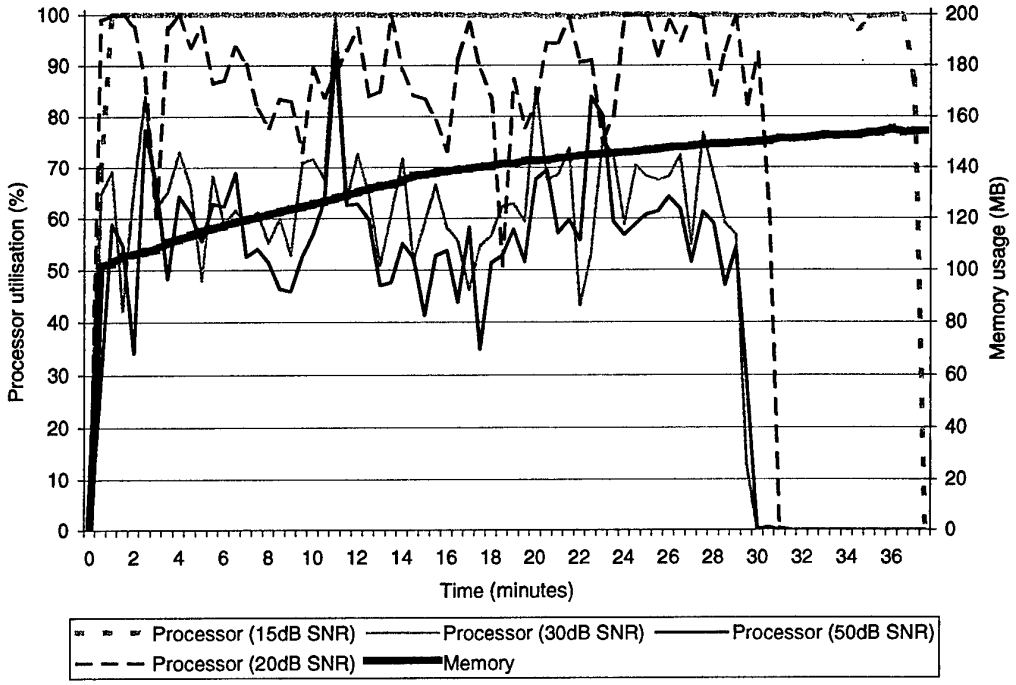

Figure 4 Graph of processor utilisation and memory usage over time while transcribing Speaker-3's 15dB, 20dB, 30dB and 50dB SNR mixed audio files with Speaker-3's 50dB SNR user profile loaded.

After the ASR loaded the Speaker-3 user profile, the memory usage was about 100MB. The memory usages while testing each of the mixed audio files were nearly identical, and increased linearly at a rate of about 1.5MB a minute until the transcription was complete.

The SNR of the audio file being transcribed affects the average processor utilisation and the time taken to transcribe that audio file. The lower the SNR of the audio file being transcribed, the higher the processor utilisation. Another major factor that affects the average processor utilisation for processor bound tasks, such as transcribing audio using ASRs, is the processor speed.

4. Upgrading Computing Power and Software Version

Lea (1982), Baker et al. (1983) and Pallett (1985) have identified over 80 different factors that affect the performance of an ASR. These factors are categorised into task related factors, human factors, language factors, channel and environmental factors, algorithmic factors and performance and response factors. Later, Makhoul and Schwartz (1994) described great advances in ASR, and attributed this to a combination of better speech modelling techniques, faster search algorithms and more powerful computers. In particular Nakatsu et al. (1994) identified computer memory and processor speed as two major task related factors that affect performance. Since 1997, Dragon Systems have released versions 3.5, 4.0, 5.0 and 6.0 of NS, and they claim an improvement in performance each time [Dragon Systems, 2002]. The software version is regarded as an algorithmic factor.

Because the WRA of the transcriptions in the background noise experiment were relatively low, an experiment was conducted with the goal to see if there was an improvement in the performance of the ASR after upgrading the computer memory and processor speed and the ASR software version.

The latest improvements in speech modelling and search algorithms could be tested using the recently released NS Professional version 5.0 and its predecessor version 4.0.

The expected life span for an average personal computer is 2.5 to 3 years [Tristan, 2001]. With this in mind, two computers with an age gap of around 3 years were selected to test how ASR performance improved with increased computing power. The more recent computer had a processor speed of 1000MHz and 512MB of memory, while the older computer had a 400MHz processor with 128MB of memory.

The experiment consisted of comparing the WRA for each combination of NS version, and computer speed and amount of memory. The clean speech, i.e. no noise added, from the 10 different Speakers in the background noise experiment were used. Five minutes from each of the clean speaker audio files were used to train each version of NS on each of the old and new computer systems. The remaining 25 minutes of audio was transcribed using each of the four combinations of computer speed and NS version. Again, the transcribed texts were compared using Sclite from NIST to produce the WRA as a percent.

The WRA results using the older computer were 56.0% and 58.7% for NS version 4.0 and 5.0 respectively. Whereas, WRA results using the new computer were 64.8% and 69.4% for NS versions 4.0 and 5.0 respectively. Hence, there is a small increase in performance between NS versions 4.0 and 5.0 (2.7% and 4.6%) and a moderate increase between old and new computer computers (8.8% and 10.7%).

ANOVA was performed studying the influence of the computer speed and amount of memory, the NS version and their interaction on the WRA. Both the computer speed and amount of memory, and the NS version were found to be significant factors. A detailed explanation of the statistical analysis is provided in Appendix B. The significant results indicate that the computer speed and amount of memory, and NS version are two independent factors that affect the WRA of the ASR. Furthermore, the WRA increases when computer speed and amount of memory or NS version are upgraded.

5. Discussion

The goals of this research were two fold: to determine whether the performance of NS achieved the best result when the levels of background noise during the training and test phases were the same, and to establish how the performance is affected when the levels of background noise during the training and test phases were different. The results of the experiment showed that there were two significant independent factors that affected the performance of NS. The most significant was the SNR in the test condition that was logarithmic in nature, where the WRA degraded moderately from 50dB to 30dB, and dramatically from 30dB to 15dB. The second was the difference between the training and test conditions that has a small negative linear relationship to the performance.

As a consequence of these factors come two valuable findings. Firstly, irrespective of the training condition, the test condition with the least amount of background noise (i.e. highest SNR) achieved the best performance. Secondly, the best performance for a particular test condition was achieved when there was no difference between training and test conditions.

Variation in the performance between different speakers was found to be another important factor. The speaker with the best performance in WRA was a senior military officer well practiced in public speaking who was comfortable to perform the recording. The speaker with the worst performance was a soldier, with orders to participate in the experiment, who was not accustomed to reading aloud, was uncomfortable and felt stressed during the recording. This variation can be attributed to inherent differences between the speakers [Doddington et al., 1998; Rabiner and Juang, 1998], their rate of speaking, the degree of articulation [Lea, 1982], and the level of stress the speaker experienced while making the recordings, which is known as the Lombard effect [Bou-Ghazale and Hansen, 2000].

Estes (1956) and Myung, Kim and Pitt (2000) suggest problems resulting from the combined effects of the arithmetic averaging of data generated from a non-linear model in the presence of individual differences. The population regression equation described in equation 3.1 was produced using arithmetic means, is non-linear and there was significant variation between individuals. However, on closer examination

the model described herein does not meet all of the conditions required for problems to occur. A detailed explanation is described in Appendix C.

The time taken for NS to transcribe some of the audio recordings exceeded their duration. An investigation into the cause of this time lag led to the analysis of the effects the SNR has on the computer's processor utilisation and memory usage over time. It was found that as the SNR of the audio recording being transcribed decreased (i.e. increase in background noise) the average processor utilisation increased. In cases where the average processor utilisation was very high, the time taken to transcribe the audio recording exceeded the 30 minute duration of the recording by as much as 7 minutes. Variation in the processor utilisation throughout the recording may be related to the variation in the rate of speech. Future work analysing these factors in more detail will be required to verify the findings presented here.

Upgrading the computer processor and memory provided a significant increase in performance. In the experiment upgrading from a computer with a 400MHz processor and 128MB of memory to a computer with a 1000MHz processor and 512MB of memory improved the WRA by an average of about 10%. Additionally, upgrading speech recognition software versions also gave a significant increase in performance, where upgrading from NS version 4.0 to 5.0 improved the WRA by an average of 4%. In one case the combined effects of upgrading the computer's processor and memory as well as NS produced an increase in WRA by 18%.

As discussed above, the SNR of the audio recording being transcribed affects the computer's average processor utilisation and the WRA of the transcription. Furthermore, the computer's processor speed affects the WRA of the transcription. Future work could include determining the relationship between the average processor utilisation and the WRA. This would provide a more complete picture of the processes involved in transcribing audio recordings using NS.

Expectedly, few insertion errors (2%) were made because recorded materials were used throughout the experiment. Deletion or rejection errors were made more frequent (38%) due to the fast speech rate preventing the ASR from being able to register utterances. This was exacerbated by use of a slow personal computer with limited memory. The substitution errors were most frequent (60%) due to the inherent inadequacies of NS speech recogniser related to its speech models and search algorithms.

There are a number of factors that may affect the accuracy of the experimental results. NS has a set of predefined commands available to allow users to make corrections, insert grammar and control aspects of software programs. Some words in the audio files were transcribed by NS as commands rather than as literal dictation. As a result some words were transcribed into punctuation, a case change, or a command to insert a new-line. New-lines and punctuations were deleted when the hypothesis text files were converted to transcript format for the scoring program Sclite. On average, 50

words (out of about 4600 words) per file were affected. This led to an increase in deletion errors of about 1.1% at the expense of substitution errors. In some cases it was difficult to identify synchronization points (beginning and end of utterance records) between the reference and hypothesis text files. In particular, the hypothesis files generated after using NS to transcribe audio files with SNRs of 15dB or 20dB. For these hypothesis text files, 8 words (out of about 4600 words) per file were affected providing an error of about 0.2%. Because the reference text files were transcribed by hand, some mistakes may have been made due to human error. These error patterns are assumed to contribute consistently across all tests, therefore no adjustments to the results were made.

6. Conclusion

This study revealed that there are two independent factors that affect the performance of Dragon NaturallySpeaking in noisy environments: the difference between the level of background noise in the training and test environments and the level of background noise in the test environment. The significance of these factors leads to the following conclusion. Regardless of the training environment, testing with the least amount of background noise achieves the best performance in terms of word recognition accuracy. However, for a particular test environment, the best performance is achieved when the levels of background noise during training and testing were the same. If the signal-to-noise ratio for the test environment is greater than 30 decibels, there should be little degradation in performance of Dragon NaturallySpeaking due to noise.

Many other factors were observed to affect the performance of Dragon NaturallySpeaking, such as the inherent difference between speakers, the rate of speech, the degree of enunciation of speech and the level of stress the speaker experienced. Increasing the speed of the computer processor and amount of physical memory produced a significant increase in performance. Additionally, upgrading Dragon NaturallySpeaking also increased performance.

7. Acknowledgements

We are very grateful for the hours of tuition and advice on Statistics given by John Hansen, Dr Glen Smith and Dr Judy Barrett from the Human Systems Integration Group at the DSTO.

Our thanks also go to Dr Kutluyil Doğançay from the University of South Australia for refereeing this report.

We would like to thank Cheryl Pope and Dr Barry Dwyer from the Computer Science Department of the University of Adelaide. Cheryl's support was essential in setting up and supervising the initial background noise experiment conducted at the University.

Barry's thoughts and advice have been much appreciated throughout the development of this report.

Many thanks go to Ashley Cooke from Theatre Operations Analysis Group at the DSTO for generously giving his time in answering countless questions.

We appreciate the time and effort made by Sean Cowan, a student from the University of Adelaide, for conducting part of the experiment with the first author. We would also like to acknowledge William Brodie-Tyrrell for writing the programs used in the experiment.

We would also like to thank Jonathan Fiscus from the US National Institute of Standards and Technology, the author of Sclite, for his valuable help and advice.

8. References

Alwang, G. (1999) *Dragon NaturallySpeaking Preferred 4.0*, PC Magazine, Nov. 5 1999, URL- http://www.zdnet.com/pcmag/

Atal, B. (1994) *Speech Technology in 2001: New Research Directions*, in Voice communication between humans and machines, ed. by Roe, D. and Wilpon, J., National Academy Press, Washington D.C., pp. 467-481.

Baker, J., Pallett, D. and Bridle, J. (1983) Speech recognition performance assessment and available databases, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, 1983, pp. 527-530.

Bou-Ghazale, E. and Hansen, J. (2000) A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress, *IEEE Transactions on Speech and Audio Processing*, **8**(4), pp. 429-442.

Clarke, A. (1997) *3001: The Final Odyssey*, HaperCollins Publishers.

Cohen, J. and Cohen, P. (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences*, Second Edition, Lawrence Erlbaum Associates, New Jersey, pp. 145-152.

Cohen, P. and Oviatt, S. (1994) *The Role of Voice in Human-Machine Communication*, in Voice communication between humans and machines, ed. by Roe, D. and Wilpon, J., National Academy Press, Washington D.C., pp. 34-75.

Doddington, G., Liggett, W., Martin, A., Przybocki, M. and Reynolds, D. (1998) SHEEP, GOATS, LAMBS and WOLVES, A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation, *International Conference on Spoken Language Processing*, Sydney, 1998, pp. 608-611.

Dragon Systems (2002), *What's New in Dragon NaturallySpeaking® Version 6?* URL – http://www.dragonsys.com/naturallyspeaking/whatsnew/.

Estes, W. (1956) The Problem of Inference from Curves Based on Group Data, *Psychological Bulletin*, **53**(2), pp. 134-140.

Fiscus, J. (1998) *Sclite Scoring Package Version 1.5*, US National Institute of Standards and Technology (NIST), URL - http://www.itl.nist.gov/iaui/894.01/tools/.

Gong, Y. (1995) Speech recognition in noisy environments: A survey, *Speech Communication*, **16**, pp. 261-291.

Juang, B. (1991) Speech recognition in adverse environments, *Computers, Speech and Language*, **5**, pp. 275-294.

Kemp, T., Schmidt, M., Westphal, M. and Waibel, A. (2000) Strategies for Automatic Segmentation of Audio Data, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, 2000, pp. 980-983.

Keppel, G. (1982) *Design and Analysis: A Researcher's Handbook*, Second Edition, Prentice-Hall, New Jersey, pp. 128-133.

Kosbar, K. (1998) *Measuring the Power of a Signal*, University of Missouri –Rolla, Electrical Engineering Dept. web page.
URL – http://www.siglab.ece.umr.edu/ee301/dsp/basics/pow.html

Lathi, B. (1998) *Modern Digital and Analog Communication Systems*, Third Edition, Oxford University Press, Inc, pp. 273.

Lea, W. (1982) What causes speech recognisers to make mistakes?, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, 1982, pp. 2030-2033.

Lea, W. (1983) Selecting the best speech recogniser for the job, *Speech Technology*, **1**(4), pp. 10-29.

Lippmann, R. (1997) Speech recognition by machines and humans, *Speech Communication*, **22**, pp. 1-15.

Makhoul, J. and Schwartz, R. (1994) *State of the Art in Continuous Speech Recognition*, in Voice communication between humans and machines, ed. by Roe, D. and Wilpon, J., National Academy Press, Washington D.C., pp. 165-198.

Mammone, R. and Zhang, X. (1998) *Robust Speech Processing as an Inverse Problem*, in The digital signal processing handbook, ed. by Madisetti, V., Williams, D., CRC Press, Florida, pp. 27.1-27.8.

Moore, D. and McCabe, G. (1993) *Introduction to the Practice of Statistics*, Second Edition, W.H. Freeman and Company, New York.

Myung, I., Kim, C. and Pitt, M. (2000) Toward an Explanation of the Power Law Artifact: Insights from Response Surface Analysis, *Memory and Cognition*, **28**(5), pp. 832-840.

Nakatsu, R. and Yoshitake, S. (1994) *What Does Voise-Processing Technology Support Today?*, in Voice communication between humans and machines, ed. by Roe, D. and Wilpon, J., National Academy Press, Washington D.C., pp. 390-421.

Nave, R. (2000) Sound Measurement in dBA, Georgia State University Physics Dept. web page.
URL - http://hyperphysics.phy-astr.gsu.edu/hbase/sound/acont.html.

Oberteuffer, J. (1994) *Commercial Applications of Speech Interface Technology: An Industry at the Threshold*, in Voice communication between humans and machines, ed. by Roe, R. and Wilpon, J., National Academy Press, Washington D.C, pp. 347-356.

Pallett, D (1985) Performance assessment of automatic speech recognisers, *Journal of Research of the National Bureau of Standards (USA)*, **90**(5), pp. 1-17.

Plutchik, A. (2000) *Dragon Naturally Speaking: A Software Review*, Sarasota PC Monitor, URL – http://www.spcug.org/reviews/0011_04.htm.

Tristan, G. (2001) *Life Cycle of Old Computers*, United States Environmental Protection Agency website, URL – http://www.epa.gov/region2/r3/problem.htm.

Weinstein, C. (1994) *Military and Government Applications of Human-Machine Communication by Voice*, in Voice communication between humans and machines, ed. by Roe, D. and Wilpon, J., National Academy Press, Washington D.C., pp. 357-370.

Appendix A: Statistical Analysis for Background Noise Experimental Results

A.1. Description of the Statistical Analysis

The influence of a factor on a continuous variable can be studied using ANOVA. In ANOVA there can be more than one factor and these factors can interact with each other. Factorial ANOVA involves equating an observed continuous variable with a linear combination of a number of factors [Keppel, 1982]. ANOVA assumes normal distribution of the sampled data, and hence the performance measure was examined with a frequency histogram across all subjects and conditions. The histogram was visually judged to be acceptable for normality and no transformation of the continuous variable was undertaken for this reason [Cohen and Cohen, 1983]. The histogram is shown in figure A.1. The two main factors suitable for ANOVA testing are the Test SNR and the Train SNR. The levels of the Test SNR and Train SNR were both five, one for each SNR of 15dB, 20dB, 30dB, 40dB and 50dB.

A (Test SNR (5) * Train SNR (5)) two-factor repeated measures ANOVA was conducted to test differences between the means for significance. Two contrasts for each factor produced a set of planned comparisons within the ANOVA. When making planned comparisons it is common practice to ignore any increase in the family wise error associated with these tests, therefore the Alpha significance level was set at 0.05% with no adjustment to the family wise error rate [Keppel, 1982].

The ANOVA F test gives a general indication whether the differences among the observed categorical means are significant [Moore and McCabe, 1993]. The omnibus F for the interaction effect was significant (F Statistic $F(16,144) = 8.01$, Mean Square Error $MSE = 3.14$, Probability $P = 0.00$) with two of the significant planned comparisons accounting for 78.72% of the interaction effect. A table of means and descriptive statistics are presented in figure A.2 and figure A.3 respectively.

The trend analysis for the interaction effect suggests that the noise level at which participants trained the ASR affects its WRA when tested at different levels of noise. However, this pattern of results can be explained more easily as the effect of the absolute difference between the Train SNR and the Test SNR rather than purely the Train SNR effect. When the Train SNR is treated as the absolute difference from the Test SNR condition, i.e. $SNR\ Difference = |Train\ SNR - Test\ SNR|$, the pattern of results becomes clearer suggesting two independent effects of the training (figure A.4; SNR Difference: $F(4,36) = 12.38, MSE = 4.89, P = 0.00$) and test (figure A.4; Test SNR: $F(4,36) = 55.89, MSE = 42.59, P = 0.00$) conditions. The Population Regression Equation below, accounts for 90.8%(figure A.5) of the cell mean performance data.

$$WRA = 5.65 + 31.22 \log_{10}(Test\ SNR) - 0.13\ SNR\ Difference$$

A.2. Frequency Histogram for the WRA

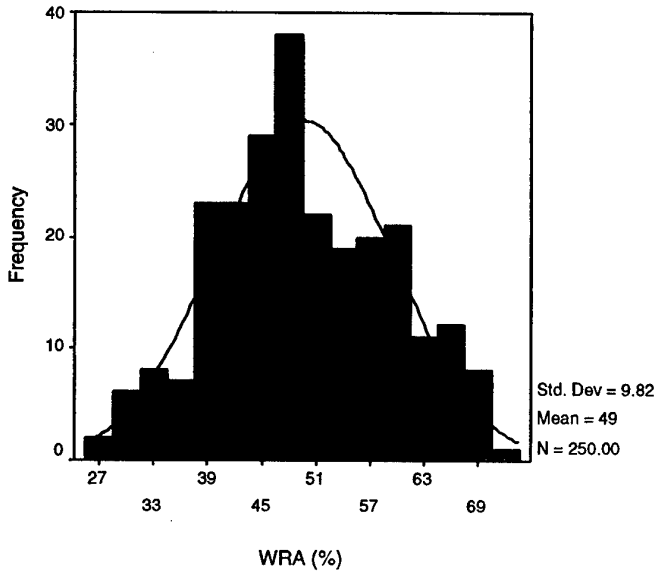

Figure A.1 Frequency Histogram of WRA (%) over all Test and Training conditions was used to determine whether the sampled data had a normal distribution. In the graph above, the line overlapping the histogram is the normal curve and, by inspection, the sampled data was judged to be suitably close to a normal curve and hence it is implied to have a normal distribution.

A.3. Arithmetic Means and Standard Errors

Trained SNR	Test SNR				
	15dB	20dB	30dB	40dB	50dB
15dB	40.3 (6.8)	46.5 (6.6)	50.8 (7.6)	52.4 (9.1)	52.8 (8.6)
20dB	40.4 (6.5)	46.8 (7.3)	51.6 (7.6)	52.7 (8.8)	53.5 (8.8)
30dB	36.9 (7.0)	47.5 (7.4)	53.8 (7.7)	55.7 (8.7)	55.2 (8.8)
40dB	36.5 (6.5)	45.1 (6.5)	51.8 (7.9)	54.8 (8.7)	55.0 (8.8)
50dB	36.0 (5.6)	45.9 (5.8)	53.0 (6.9)	55.0 (8.6)	56.0 (8.6)

Figure A.2 Table of arithmetic means and standard errors (in parentheses) of WRA for different Training and Test SNRs across ten Speakers.

A.4. Test of Within-Subject Effects for Train SNR and Test SNR

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Train SNR	Sphericity Assumed	61.370	4	15.342	1.812	***.148***
Error(Train SNR)	Sphericity Assumed	304.846	36	8.468		
Test SNR	Sphericity Assumed	9945.234	4	2486.309	57.1	***.000***
Error(Test SNR)	Sphericity Assumed	1567.886	36	43.552		
Train SNR * Test SNR	Sphericity Assumed	480.258	***16***	30.016	***8.735***	***.000***
Error(Train SNR * Test SNR)	Sphericity Assumed	494.814	***144***	***3.436***		

*Figure A.3 Table of results for within subject effects using factors: Training SNR, Test SNR and their interaction, Training SNR * Test SNR. A factor has a significant component when the Sig. value is near zero. The F statistic for the interaction effect was significant ($F(16,144)=8.74$, Mean Square Error $MSE=3.44$, Probability Value $P=0.00$).*

A.5. Test of Within-Subject Effects for SNR Difference and Test SNR

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Test SNR	Sphericity Assumed	384.766	***4***	96.191	***19.858***	***.000***
Error(Test SNR)	Sphericity Assumed	174.386	***36***	***4.844***		
SNR Diff.	Sphericity Assumed	9945.234	***4***	2486.31	***57.088***	***.000***
Error(SNR Diff.)	Sphericity Assumed	1567.886	***36***	***43.552***		
Test SNR * SNR Diff.	Sphericity Assumed	156.862	***16***	9.804	***2.258***	***.059***
Error(Test SNR*SNR Diff.)	Sphericity Assumed	625.274	***144***	***4.342***		

*Figure A.4 Table of results for within subject effects using factors: SNR Difference, Test SNR and their interaction, SNR Difference * Test SNR. A factor has a significant component when the Sig. value is near zero. The SNR Difference ($F(4,36)=17.20$, $MSE=4.22$, $P=0.00$) and the Test SNR ($F(4,36)=55.89$, $MSE=42.59$, $P=0.00$) are both significant, while their interaction ($F(16,144)=2.37$, $MSE=5.02$, $P=0.06$) is not.*

A.6. Regression Model

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SNR Difference, Log(Test SNR)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: WRA

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.953[a]	***.908***	.899	2.0615

a. Predictors: (Constant), SNR Difference, Log(Test SNR)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	920.251	2	460.126	108.268	.000[a]
	Residual	93.497	22	4.250		
	Total	1013.749	24			

a. Predictors: (Constant), SNR Difference, Log(Test SNR)

b. Dependent Variable: WRA

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	***5.645***	3.174		1.778	.089
	Log(Test SNR)	***31.224***	2.158	.938	14.466	.000
	SNR Difference	***-.133***	.038	-.229	-3.534	.002

a. Dependent Variable: WRA

Figure A.5 Factors entered for a Multiple Regression Model were WRA (the dependant variable), SNR Difference and the Logarithm to the base 10 of the Test SNR (the independent variables). The Population Regression Equation, which was formulated using these factors and their corresponding Regression Coefficients, follows: $WRA(\%) = 5.65 + 31.22\log_{10}(Test\ SNR) - 0.13\,SNR\ Difference$. *The R Square, also known as the Squared Multiple Correlation, is 0.908 or 90.8%. This R Square value estimates how well the model fits the sampled data. Values close to one or 100% indicate that the model fits the data well [Moore and McCabe, 1993].*

Appendix B: Statistical Analysis for Computing Power and Software Version Experimental Results

B.1. Description of the Statistical Analysis

ANOVA was used to determine the influence of Computing Power, and NS Version on the WRA. The two main factors suitable for ANOVA testing were the Computing Power and the NS Version. The levels of the Computing Power and NS Version were both two.

A (Computing Power (2) * NS Version (2)) two-factor repeated measures ANOVA was conducted to test differences between the means for significance. The F statistic for the Computing Power (F *Statistic* $F(1,9) = 252.35$, *Mean Square Error* $MSE = 4.08$, *Probability Value* $P = 0.00$) and NS Version ($F(1,9) = 507.65$, $MSE = 0.31$, $P = 0.00$) were both found to be significant. However, the interaction effect ($F(1,9) = 2.53$, $MSE = 1.45$, $P = 0.15$) was not significant. A table of means and descriptive statistics are presented in figure B.1 and B.2 respectively.

B.2. Arithmetic Means and Standard Errors

Computing Power: Speed/Memory	NS Version	
	4.0	5.0
400MHz/128MB	56.0 (8.6)	58.7 (8.1)
1000MHz/512MB	64.8 (9.4)	69.4 (10.2)

Figure B.1 Table of arithmetic means and standard errors, in parentheses, of WRA for different Computing Power and NS Version combinations across ten Speakers.

B.3. Test of Within-Subject Effects for Computing Power and NS Software Version

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Computing Power	Sphericity Assumed	.567	***1***	.567	***119.52***	***.000***
Error(Computing Power)	Sphericity Assumed	4.267E-02	***9***	***4.7E-03***		
NS Version	Sphericity Assumed	7.466E-02	***1***	7.5E-02	***132.38***	***.000***
Error(NS Version)	Sphericity Assumed	5.076E-03	***9***	***5.6E-04***		
Computing Power*NS Version)	Sphericity Assumed	7.969E-03	***1***	8.0E-03	***10.133***	***.011***
Error(Computing Power*NS Version)	Sphericity Assumed	7.078E-03	***9***	***7.9E-04***		

*Figure B.2 Table of results for within subject effects using factors: Computing Power, NS Version and their interaction, Computing Power * NS Version. A factor has a significant component when the Sig. value is near zero (i.e. Sig.* ≤ 0.05*). The Computing Power (*$F(1,9) = 252.35$, $MSE = 4.08$, $P = 0.00$*), the NS Version (*$F(1,9) = 507.65$, $MSE = 0.31$, $P = 0.00$*) are both significant, while their interaction (*$F(1,9) = 2.53$, $MSE = 1.45$, $P = 0.15$*) is not significant.*

Appendix C:
Making inferences from curves based on group data

Estes (1956) warns against making inferences from curves based on group data. Myung, Kim and Pitt (2000) also suggest problems resulting from the combined effects of arithmetic averaging of data generated from a non-linear model in the presence of individual differences.

Recall the population regression equation below.

$$WRA = 5.65 + 31.22 \log_{10}(Test\ SNR) - 0.13\ SNR\ Difference$$

This equation is of the form $y = a + b \log x_1 - c x_2$, where y is the dependent variable, x_1 and x_2 are independent variables, and a, b and c are constant parameters. There are three important points to consider regarding this model. Firstly, this equation was derived using the arithmetic means of the Word accuracies across the ten speakers. Secondly, the relationship between y and x_1 is non-linear. Thirdly, there is a significant variation in WRA scores between individual speakers.

Estes (1956) defines three classes of functions, A, B and C that should be treated differently. Class A functions are unmodified by averaging and an equation of the form $y = a + b \log x_1 - c x_2$ is categorised as a class A function. Class A functions are unmodified by averaging because each parameter in the function either stands alone or is a coefficient multiplying another function which depends only on the independent variable. When averaging this case the function that depends only on the independent variable factors out. This leaves the mean value of the parameter multiplying the function that depends only on the independent variable. Equation C.1 illustrates this with n individual subjects for the model of the form $y = a + b \log x_1 - c x_2$, where $\bar{a}$, $\bar{b}$ and $\bar{c}$ are the arithmetic means of the parameters a, b and c for models based on the individual results. For instance, the model for subject 1 is $y = a_1 + b_1 \log x_1 - c_1 x_2$.

$$\begin{aligned} y &= \frac{1}{n}\sum_{i=1}^{n} a_i + b_i \log x_1 - c_i x_2 \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n} b_i\right)\log x_1 - \left(\frac{1}{n}\sum_{i=1}^{n} c_i\right)x_2 \\ &= \bar{a} + \bar{b} \log x_1 - \bar{c} x_2 \end{aligned} \qquad (C.1)$$

Myung et al. (2000) describe a non-linear model as a model where the dependent variable y is non-linearly related to the parameters b and c rather than the independent variables x_1 and x_2. They consider the power and exponential models to be non-linear, while linear and logarithmic models to be linear.

DISTRIBUTION LIST

The Effects of Background Noise on the Performance of an Automatic Speech Recogniser

Jason Littlefield and Ahmad Hashemi-Sakhtsari

AUSTRALIA

DEFENCE ORGANISATION

Task Sponsor
Comd DJFHQ
HKS

S&T Program
Chief Defence Scientist
FAS Science Policy
AS Science Corporate Management
Director General Science Policy Development
} shared copy
Counsellor Defence Science, London (Doc Data Sheet)
Counsellor Defence Science, Washington (Doc Data Sheet)
Scientific Adviser to MRDC Thailand (Doc Data Sheet)
Scientific Adviser Joint
Navy Scientific Adviser (Doc Data Sheet and distribution list only)
Scientific Adviser - Army
Air Force Scientific Adviser
Director Trials

Information Sciences Laboratory
Ian Heron, Chief of Command and Control Division
Rudi Vernik, Research Leader, Command and Intelligence Environments Branch
Dale Lambert, Head, Human Systems Integration Group
Ahmad Hashemi-Sakhtsari, Task Manager, Human Systems Integration Group
Alex Yates, Theatre Operations Analysis
Ashley Cooke, Theatre Operations Analysis
Author(s):
Jason Littlefield, Human Systems Integration Group
Ahmad Hashemi-Sakhtsari, Human Systems Integration Group

DSTO Library and Archives
Library Edinburgh 2 copies
Australian Archives
Library Canberra

Capability Systems Staff
Director General Maritime Development (Doc Data Sheet only)
Director General Land Development
Director General Aerospace Development (Doc Data Sheet only)

Knowledge Staff

Director General Command, Control, Communications and Computers (DGC4) (Doc Data Sheet only)

Army

Chief of Staff, DJFHQ, Enoggera, Qld 4051
DC2I, Knowledge Systems
ABCA National Standardisation Officer, Land Warfare Development Sector, Puckapunyal (4 copies)
SO (Science), Deployable Joint Force Headquarters (DJFHQ) (L), Enoggera QLD
NPOC QWG Engineer NBCD Combat Development Wing, Puckapunyal, VIC

Intelligence Program

DGSTA Defence Intelligence Organisation
Manager, Information Centre, Defence Intelligence Organisation

Defence Libraries

Library Manager, DLS-Canberra
Library Manager, DLS - Sydney West (Doc Data Sheet Only)

UNIVERSITIES AND COLLEGES

Australian Defence Force Academy
Library
Head of Aerospace and Mechanical Engineering
Serials Section (M list), Deakin University Library, Geelong, VIC
Hargrave Library, Monash University (Doc Data Sheet only)
Librarian, Flinders University
Barry Dwyer, Department of Computer Science, The University of Adelaide, North Terrace, SA
Cheryl Pope, Department of Computer Science, The University of Adelaide, North Terrace, SA
Kutluyil Doğançay, School of Electrical and Information Engineering, The University of South Australia, Mawson Lakes, SA

OTHER ORGANISATIONS

National Library of Australia
NASA (Canberra)
AusInfo
State Library of South Australia

OUTSIDE AUSTRALIA

INTERNATIONAL DEFENCE INFORMATION CENTRES

US Defense Technical Information Center, 2 copies
UK Defence Research Information Centre, 2 copies
Canada Defence Scientific Information Service, 1 copy
NZ Defence Information Centre, 1 copy

ABSTRACTING AND INFORMATION ORGANISATIONS

Library, Chemical Abstracts Reference Service

Engineering Societies Library, US
Materials Information, Cambridge Scientific Abstracts, US
Documents Librarian, The Center for Research Libraries, US

INFORMATION EXCHANGE AGREEMENT PARTNERS

Acquisitions Unit, Science Reference and Information Service, UK
Library - Exchange Desk, National Institute of Standards and Technology, US

SPARES (5 copies)

Total number of copies: 59

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

The Effects of Background Noise on the Performance of an Automatic Speech Recogniser

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document (U)
Title (U)
Abstract (U)

4. AUTHOR(S)

Jason Littlefield and Ahmad Hashemi-Sakhtsari

5. CORPORATE AUTHOR

Information Sciences Laboratory
PO Box 1500
Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-RR-0248	AR-012-500	Research Report	November 2002

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
9505-21-138	JTW 01/092	HKS & Comd. DJFHQ	25	31

13. URL on the World Wide Web

http://www.dsto.defence.gov.au/corporate/reports/DSTO-RR-0248.pdf

14. RELEASE AUTHORITY

Chief, Command and Control Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DEFTEST DESCRIPTORS

Speech recognition
Speech analysis
Noise effects

19. ABSTRACT

Ambient or environmental noise is a major factor that affects the performance of an automatic speech recogniser. Large vocabulary, speaker-dependent, continuous speech recognisers are commercially available. Speech recognisers perform well in a quiet environment, but poorly in a noisy environment. Speaker-dependent speech recognisers require training prior to them being tested, where the level of background noise in both phases affects the performance of the recogniser. This study aims to determine whether the best performance of a speech recogniser occurs when the levels of background noise during the training and test phases are the same, and how the performance is affected when the levels of background noise during the training and test phases are different. The relationship between the performance of the speech recogniser and upgrading the computer speed and amount of memory as well as software version was also investigated.